Central reassessment of GH concentrations measured at local treatment centers in children with impaired growth: consequences for patient management

Berthold P Hauffa 1, Nils Lehmann 2, Markus Bettendorf 3, Otto Mehls 3, Helmut-Günther Dörr 4, Carl-Joachim Partsch 5, Hans P Schwarz 6, Nikolaus Stahnke 7, Heinz Steinkamp 8, Elfriede Said 8, Sabine Sander 8 and Michael B Ranke 9 and participating members of the German KABI International Growth Study/Internationale Genotropin-Langzeituntersuchung (KIGS/IGLU) Study Group

1Department of Pediatric Hematology/Oncology and Endocrinology, University Children’s Hospital, Essen, Germany, 2Institute for Medical Informatics, Biometry and Epidemiology, University of Essen, Essen, Germany, 3University Children’s Hospital, Heidelberg, Germany, 4University Hospital for Children and Adolescents, Erlangen, Germany, 5Pediatric Endocrinology, University Children’s Hospital, Kiel, Germany, 6University Children’s Hospital, München, Germany, 7University Children’s Hospital, Hamburg, Germany, 8Pharmacia Corporation, Erlangen, Germany and 9Pediatric Endocrinology Section, University Children’s Hospital, Tuebingen, Germany

(Correspondence should be addressed to B P Hauffa, Klinik für pädiatrische Hämatologie/Onkologie und Endokrinologie, Zentrum für Kinderheilkunde und Jugendmedizin, Universitätsklinikum Essen, Hufelandstr. 55, D-45122 Essen, Germany; Email: berthold.hauffa@uni-essen.de)

Abstract

Objective: GH deficiency is diagnosed in children if serum GH fails to rise above a predefined cutoff value in response to at least two stimuli. Diagnostic decisions based on this testing are highly variable between centers and depend on the GH assays used. Considering the large spectrum of commercially available GH assays, we wanted to evaluate the agreement between assays, and to test whether assay-related variability of diagnostic decisions could be reduced by reassessment of peak GH concentrations in a reference center.

Design: We reanalysed 699 peak GH serum samples obtained after GH testing of 382 children and adolescents from 19 centers using three reference assays and compared these results with those obtained with the local assays. A subgroup of 132 patients tested with the combination of insulin hypoglycemia test and arginine test was evaluated for changes in the assignment to the diagnostic group of GH deficiency.

Results: The mean difference between methods ranged from 5.4 to 10.3 mU/l, slopes of the regression lines from 1.28 to 1.65. Significant non-linearity was detected in five of six assay comparisons, indicating that most assay results cannot be interconverted by the use of a factor. Overall agreement between reference and local assays was only moderate. Significant changes in diagnostic assignment occurred when different assays were used on the same patient (P = 0.0001 – P = 0.0023). Based on GH remeasurement by one reference assay, 36 of 132 patients were categorized differently, with 35 patients changing into the GH-deficient group. Similar findings were obtained with the other reference assays.

Conclusions: To decrease variability in GH testing related to assays and cutoff values, we recommend nationwide reassessment of GH peak sera in reference centers. Decisions to treat GH deficiency should incorporate the reference center results.

Introduction

In most countries, the diagnosis of growth hormone (GH) deficiency in slowly growing children is based on the failure of serum GH concentrations to rise above a predefined cutoff value in response to at least two pharmacologic or physiologic stimuli (1). Considerable variability becomes apparent when patients diagnosed as GH deficient are retested (2, 3).

To the intrinsic variability of test results caused by physiological influences on the regulation of GH release (4–6), variability is added by the testing instruments themselves. This includes the many different combinations of pharmacologic stimuli used for testing the somatotropic axis (7), and the multitude of commercially available assays for measuring GH. As a consequence, recognizing a child with GH deficiency depends largely on the GH assay used in a treatment center, and its validation for a specific setting of test procedures and cutoff values (8).

To further explore assay-related variability, we wanted to evaluate whether the results of GH testing in daily practise could be made comparable between centers. We tested the agreement between the GH
peak concentrations measured in the participating centers with their local assays and the results obtained at repeat measurement in one reference center using three different GH assays. In addition, we asked how the diagnostic assignment would change when different assays were employed on sera from the same patient.

Materials and methods

Samples and subjects
Twenty-one German secondary and tertiary care centers diagnosing and treating children with growth disorders participated in the study. From 1 January 1998 to 31 December 1999 all consecutive children and adolescents who underwent functional testing of the somatotropic axis or evaluation of spontaneous GH secretion were considered for inclusion. GH concentrations were measured first using the local GH assay in each center. If serum was available from the samples containing the peak GH concentration, this was sent to a central laboratory (University Children’s Hospital, Tuebingen) for re-evaluation. Seven hundred and ninety serum samples from 434 patients were reassayed centrally. Ninety-one of these results (52 patients, two centers) were excluded from the analysis. In 32 patients (61 samples), locally obtained GH results were not communicated. In nine patients (14 samples) the units used for reporting the GH data could not be verified. In seven patients (seven samples) the local assay was not reported, in three patients (six samples) the GH testing procedure remained unclear. In one patient who had undergone arginine testing twice, only one set of data was used. On using GH-releasing hormone stimulation, one patient showed low GH values when measured with the immunofunctional assay (IFA), in contrast to very high GH values obtained with all other assays, indicating the presence of a GH moiety with lesser bioactivity; this patient (two samples) was also omitted from the analysis. Thus, 699 peak GH serum samples of 382 children and adolescents (258 male, 124 female) from 19 centers were available for data analysis.

A subgroup of 132 patients with the most abundant test combination (insulin hypoglycemia test and arginine test) was evaluated for changes in the assignment to the diagnostic group of GH deficiency. Patients were diagnosed as GH deficient if the GH concentration remained below the assay-specific cutoff value in both tests. Otherwise, the patients were categorized as not GH deficient.

Ethical approval
The study protocol was reviewed and approved by the Ethics Committee for Medical Research of the University of Tuebingen. Patients were only submitted for functional testing of the somatotropic axis when results of the clinical examination or other laboratory tests indicated the presence of GH deficiency. Since some of the functional tests (e.g., insulin hypoglycemia) imply some risk, enough blood is routinely withdrawn for repeat laboratory measurements should anything go wrong in the laboratory. If the patient serum was not used up for that purpose, the remainder was sent to the reference laboratory for further measurements within our study. No additional blood was withdrawn for this study. Therefore the Institutional Review Board of the University of Tuebingen has not required us to obtain written informed consent from the parents for further analysis of the remaining serum.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Characteristics and frequency of use of different human (h) GH immunoassay methods employed by the participating centers.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assay no.</td>
<td>Assay name</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>1</td>
<td>HGH MAIAclone</td>
</tr>
<tr>
<td>2</td>
<td>SERIA hGH</td>
</tr>
<tr>
<td>3</td>
<td>Nichols hGH</td>
</tr>
<tr>
<td>4</td>
<td>DPC Immulite hGH</td>
</tr>
<tr>
<td>5</td>
<td>Varelisa</td>
</tr>
<tr>
<td>6</td>
<td>hGH-EASIA</td>
</tr>
<tr>
<td>7</td>
<td>hGH-RIA</td>
</tr>
</tbody>
</table>

IRMA, immunoradiometric assay; CLIA, chemoluminescence immunoassay; EIA, enzyme immunoassay; RIA, radioimmunoassay; 1. IRP, first international reference preparation; 1. IS, first international standard.

*Response to pharmacologic stimuli or assessment of spontaneous secretion. †As supplied by the manufacturer.
Analytical methods

The GH assays used in the participating centers (local assays) and their characteristics are detailed in Table 1. Samples containing the peak GH concentration were stored frozen for various lengths of time not exceeding 6 months at −20°C in 300 µl aliquots and sent to the central laboratory at the University Children’s Hospital, Tuebingen (M B R), where GH concentration was reassessed in duplicate using three different reference assays. GH in serum measured with the reference assays was reported to be stable under these conditions over a time-period of up to 1 year.

The first reference assay (Tuebingen in-house assay) was an RIA calibrated against the First International Standard for somatropin (88/624). Its first polyclonal antibody is directed exclusively against 22 kDa recombinant hGH. Analytical sensitivity was 0.15 mU/l. The intra-assay (interassay) coefficients of variation (CV) were ≤ 7.9% at GH concentrations ranging from 3.6 to 56.4 (8.7–30.6) mU/l. The cutoff concentration established in Tuebingen was applied for rating the peak GH response as normal (≥ 24 mU/l) or insufficient (< 24 mU/l).

The second reference assay (AutoDELFIA hGH; Wallac Oy, Turku, Finland) was an enzyme-linked immunosorbent assay (ELISA) calibrated against the First International Standard for hGH (80/505) with two monoclonal mouse antibodies directed against different epitopes of 22 kDa hGH. Analytical sensitivity was 0.03 mU/l. Cross-reactivity with other proteohormones was < 0.1%. The intra-assay (interassay) CV values at GH concentrations from 0.52 to 89.2 mU/l were ≤ 6.3% (≤ 5.0%). The cutoff value distinguishing between a normal GH reserve and GH deficiency was set at 26 mU/l. In 20 patients (26 samples), there was not enough serum for parallel determinations with this assay.

The third assay used for reference (Active Immuno-functional hGH ELISA; Diagnostic Systems Laboratories, Webster, TX, USA) was an IFA calibrated against the First International Standard for somatropin (88/624). A first monoclonal mouse antibody is directed against a bioactive epitope of 22 kDa hGH. Recombinant hGH-binding protein (GHBP) is the second binding agent. Analytical sensitivity was 0.18 mU/l. Cross-reactivity with 20 kDa pituitary hGH was 53.3%. The intra-assay (interassay) CV values at GH concentrations from 1.47 to 38.13 (1.44–45.8) mU/l were ≤ 5.1% (≤ 9.6%). According to the manufacturer, the cutoff value distinguishing between a normal GH reserve and GH deficiency was set at 15 mU/l. In nine patients (nine samples), there was not enough serum for parallel determinations with this assay.

The central laboratory (M B R) was blinded to the origin of the samples, the type of assay used and the GH values obtained at the respective centers, and the cutoff levels employed for the diagnosis of GH deficiency at these centers. The evaluating center (B P H) was blinded to the name of the referring center. Results are expressed as mU/l. Data given in ng/ml were recalculated as mU/l using the conversion factors given by the assay manufacturers (Table 1).

Statistical methods

GH results obtained at the different centers were combined and compared pairwise with the results of the central laboratory by the Bland–Altman plot (9) and the Passing–Bablok regression (10), using MedCalc software Mariakerke, Belgium (11). These methods are non-parametric and do not require one assay to be assigned the role of an analytical gold standard. Subjecting the GH data to a common Box–Cox power transformation with $y = (x^{0.4} - 1)/0.4$ yielded a normal distribution. From the transformed data, the concordance correlation coefficient (12) was calculated as a measure of agreement between assays. Values for overall concordance ($\rho$) between 0.2 and 0.7 indicate minor concordance, values between 0.7 and 0.85 indicate moderate concordance, and values larger than 0.85 indicate clear high concordance. Whether the local assays and each of the three assays used by the central laboratory agreed in the assignment of a patient to a diagnostic group (GH deficient vs not GH deficient) was assessed using the McNemar test and the $\kappa$ statistic (13). A $\kappa$ value between 0.4 and 0.6 indicates moderate agreement, $\kappa$ between 0.6 and 0.8 indicates clear agreement, and $\kappa$ between 0.8 and 1.0 indicates high agreement. The latter analyses were performed using SAS software version 8.2 (14).

Results

First, assay comparability with regard to peak GH concentrations was assessed using the Bland–Altman plot and the Passing–Bablok regression. Results obtained with the Bland–Altman method are displayed as scatterplots in Fig. 1. When the local assay results were compared with the three reference assays, the scatterplots displayed a widening range of differences with increasing GH concentrations, indicating that the variation of at least one method of each pair depended strongly on the GH concentration present in the sample. Also, because the mean of the assay differences was not zero, a bias between methods was demonstrated. This increased from the Tuebingen assay (− 5.4 mU/l) to the DELFIA assay (− 6.0 mU/l) to the IFA, which showed the greatest difference, measuring on average 10.3 mU/l lower than the local assays. The ± 1.96 s.d. range of the assay differences was large when the local assays were compared with the reference assays: DELFIA, ± 51.1 mU/l; Tuebingen assay, ± 52.2 mU/l; IFA, ± 54.5 mU/l. As the GH
cutoff concentrations distinguishing between GH deficiency and normal GH secretion were reported to be in the range of 20–26 mU/l at most centers. The implication is that apparently normal GH concentrations measured with the local assays can lie in the GH-deficient range when measured with the reference assays, and vice versa. In addition, a proportional error component was detected in the scatterplots, visible below the mean $\pm 1.96$ S.D. lines as proportionally larger negative differences. This component is due to higher readings of the local assays in the high GH concentration range. Therefore, GH measurements obtained with these assays cannot readily be interconverted by the use of some factor after correction for the bias.

A different picture emerged when the reference assays were compared with each other. The best agreement was found between the Tuebingen assay and the DELFIA, with a bias of only $-0.5$ mU/l and a $\pm 1.96$ S.D. range of $\pm 16.9$ mU/l. Comparison between the Tuebingen assay (DELFIA) and IFA revealed a slightly higher bias of $-4.7$ mU/l ($-4.3$ mU/l) and a similar $\pm 1.96$ S.D. range of $\pm 15.8$ mU/l ($\pm 20.9$ mU/l).

Additional information was gained by the Passing–Bablok analysis. Again, the least disagreement was found between the Tuebingen assay and the DELFIA.
Regression equations with 95% CI values for the intercept and the slope are shown in Table 2. The data were also tested for deviations from a simple linear relationship. A significant non-linearity was found with all comparisons except the Tuebingen–DELFIA pair. This shows that a conversion of the GH concentrations between assays based on a linear equation is not valid in these cases.

The strength of the agreement between assay results was quantified using the concordance correlation. With this method, overall concordance was highest between DELFIA and the Tuebingen and IFA assays (ρc > 0.86). The concordance was lower between the local assays and the three reference assays (ρc ≤ 0.80) and lowest for the local assays–IFA pair (ρc = 0.71).

There were consequences for these differences in GH measurement in the assignment of the patients to the diagnosis of GH deficiency. A subgroup of 132 patients, in whom the diagnosis was based on a combination of an insulin hypoglycemia test and an arginine test, was investigated further. GH deficiency was assumed if the stimulated GH concentration was below the specified cutoff value in both tests. The diagnostic matrices for the pairwise assay comparisons are shown in Fig. 1 (graphs). A significant imbalance in diagnostic assignment occurred when different assays were used on the same patient (P < 0.0001–0.0023); the only exception was the Tuebingen–DELFIA comparison. While the Tuebingen and DELFIA assays showed excellent agreement (κ = 0.88), only moderate agreement was found when local assays and reference assays were compared.

**Discussion**

Our data show that agreement between the results of GH assays used in 19 German tertiary and secondary centers experienced in the treatment of children with growth disorders is extremely variable. As a consequence, the probability that a short child will be diagnosed as GH deficient depends to a great extent on the GH assay and the cutoff GH concentrations used at the local center. These vary considerably between centers. We have also demonstrated that sending peak GH serum samples to a central laboratory for repeat GH measurement and using the remeasured data to adjust treatment decisions is a feasible way to reduce variability in the treatment decisions of childhood GH deficiency based on pharmacological testing in the setting of a nationwide health system.

More than 200 papers have addressed the subject of the variability of tests of GH release in children. This work has been extensively reviewed, and consensus guidelines have been formulated for the diagnosis of GH deficiency (1, 15) with the aim of decreasing this variability. There is still a gap, however, between these guidelines and the daily practice in European countries (16).

One major source of variability is the GH assay itself, as shown by the outcome of multilaboratory external quality-control assessment programs (17, 18). First, results depended on the assay standards. Three different International Reference Preparations or standards were used by the local and reference assays of this study. Two of them (IRP 66/217 and IS 80/505) are pituitary standards, originally defined in units (U) by bioassay, and 1 (IS 88/624) is a recombinant standard defined in mass units (mg) by amino acid analysis (19). Therefore, assay results based on the first two standards should be expressed in units. However, three of the seven study assays calibrated against pituitary standards expressed their results in ng/ml. These three assays all used the standard IS 80/505. To reconvert GH results to mU/l, one manufacturer recommended a factor of 2.0, whereas two manufacturers recommended a factor of 2.6. Thus, variability arose not only from the use of different standards but also from the way of expressing the results.

Antibodies represented a major source of variation in our assay comparisons. GH in biologic samples is essentially a mixture of more than 200 different molecular variants (20). Polyclonal antibodies raised against extractive GH will therefore bind to a multitude of molecular variants regardless of their biologic activity. At the other end of the spectrum are monoclonal antibodies directed against two sites on the 22 kDa GH necessary for receptor dimerization, and GHBP itself to detect predominantly biologically active GH variants (21). One of the reference assays (IFA) using the latter antibodies in fact yielded the lowest GH values, representing 60.6% of the concentrations measured with the local assays. 76.5% of the concentrations of the
DELFIA assay, and 88.9% of the Tuebingen assay concentrations, with the Tuebingen assay using the same recombinant standard (1. IRP 88/624). This is in accordance with previously published slopes of IFA regression lines (IFA vs the HybriTech assay 65%) (22). Further, it has been reported that in some patients GH peaks appear later after stimulation when measured with an IFA. In a study like ours, in which only the peak serum samples identified by local assays were sent for interlaboratory comparison, this could introduce a bias toward lower GH concentrations measured by IFA (23).

At first sight, some of our data seem to support the notion that assay results could be made comparable by the use of simple conversion factors. However, significant non-linearity existed in all comparisons except the Tuebingen–DELFIA pair, precluding the use of a linear conversion factor. Taking into account the known sources of GH variability, it is not surprising that putting together a group of local assays and comparing it with single reference assays yields higher variability than comparison of two single assays. This, however, reflects the situation in the health care systems of many countries, where the decision to treat a GH-deficient child is based on data from numerous different assays. In the USA, 25 commercially available GH assays are currently listed with the Food and Drug Administration (FDA report 58 FR 39680; CLIA database 12/02: http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfCLIA/search.cfm).

It has been repeatedly stressed that any GH assay and test method must be used exclusively with the cutoff GH value validated for that specific situation (1, 23). In daily practise, arbitrary cutoff GH values are frequently used. Assay variability combined with the use of arbitrary cutoff values for the diagnosis of GH deficiency leads to significant differences with regard to GH treatment decisions. In one single-center assay comparison study using fixed chosen cutoff values, correlation between individual assay results was high. Nevertheless, agreement between treatment decisions obtained with these assays was low (24). Even in our patients, where cutoff values suggested by the assay manufacturers were used, the diagnostic assignment of patients varied considerably between assays. This finding is in contrast to earlier reports, in which the use of specific cutoff values prevented major non-agreement in diagnostic assignment among some assays (25).

It is mandatory that all patients with GH deficiency have the same chance of being diagnosed and treated with GH independent of the methods used. Unnecessary treatment must be avoided, because treatment risks, albeit small, do exist (26–28), and the costs of treatment are high.

We suggest that central repeat measurement of peak GH sera in a reference center are a good way of improving comparability of GH test data within a health system. Repeat measurement could be done by one or two reference assays, perhaps by a combination of a polyclonal assay and an IFA. Since it is not ethically and financially feasible to validate cutoff values for each of the great number of GH assay–functional test combinations currently in use, limitation to small number of reference assays would enable us to set up cutoff values based on sufficient numbers of patients and healthy children with different height velocities.

In conclusion, our data discourage the use of linear conversion systems to make GH results obtained with different assays comparable and to bring diagnostic decisions based on these assays to agreement. To decrease variability related to assays and cutoff values in diagnosing GH deficiency, we recommend nationwide reassessment of GH peak serum samples of all pediatric patients in central laboratories using reference assays. Decisions to treat GH deficiency should incorporate these results.

Acknowledgements

The participating investigators of the German KIGS/IGLU Study Group are as follows: W Andrer (Datteln), R Arndt (Neu- brandenburg), M Bettendorf (Heidelberg), H J Böhles (Frankfurt/Main), H G Dörr (Erlangen), S Eisberg (Minden), B P Hauffa (Essen), O Hiort (Lübeck), E Korsch (Köln), M Mix (Rostock), K Mohnike (Magdeburg), M Morlot (Hannover), R Mühlberg (Krefeld), C-J Partsch (Kiel), M B Ranke (Tuebingen), E Schönau (Köln), H-P Schwarz (München), G Simic-Schleicher (Bremen), K-P Ullrich (Gotha), T van Meegen-Freund (Aachen), and Ch Vogel/A Klinghammer (Chemnitz). This work was supported by a grant from Pharmacia Corporation, Erlangen, Germany to B P H and M B R. The authors thank Mrs K Weber for her expert technical assistance and R D Oader, Dept. of Child and Adolescent Psychiatry, University of Essen, for helpful review of the manuscript.

References

6 Veldhuis JD & Irmananesh A. Physiological regulation of the human growth hormone (GH)–insulin-like growth factor type 1
(IGF-I) axis: predominant impact of age, obesity, gonadal function, and sleep. *Sleep* 1996 **19** S221–S224.


19 Bristow AF. International standards for growth hormone. *Hormone Research* 1999 **51** 7–12.


